

HIGH-DIMENSIONAL LONGITUDINAL CLASSIFICATION WITH THE MULTINOMIAL FUSED LASSO

BY SAMRACHANA ADHIKARI^{1,*}, FABRIZIO LECCI^{1,*}, JAMES T. BECKER², BRIAN W. JUNKER¹,
LEWIS H. KULLER², OSCAR L. LOPEZ² AND RYAN J. TIBSHIRANI¹

¹*Carnegie Mellon University*

²*University of Pittsburgh*

We study regularized estimation in high-dimensional longitudinal classification problems, using the lasso and fused lasso regularizers. The constructed coefficient estimates are piecewise constant across the time dimension in the longitudinal problem, with adaptively selected change points (break points). We present an efficient algorithm for computing such estimates, based on proximal gradient descent. We apply our proposed technique to a longitudinal data set on Alzheimer's disease from the Cardiovascular Health Study Cognition Study, and use this data set to motivate and demonstrate several practical considerations such as the selection of tuning parameters, and the assessment of model stability.

1. Introduction. In this paper, we study longitudinal classification problems in which the number of predictors can exceed the number of observations. The setup: we observe n individuals across discrete timepoints $t = 1, \dots, T$. At each timepoint we record p predictor variables per individual, and an outcome that places each individual into one of K classes. The goal is to construct a model that predicts the outcome of an individual at time $t + \Delta$, given his or her predictor measurements at time t . Since we allow for the possibility that $p > n$, regularization must be employed in order for such a predictive model (e.g., based on maximum likelihood) to be well-defined. Borrowing from the extensive literature on high-dimensional regression, we consider two well-known regularizers, each of which also has a natural place in high-dimensional longitudinal analysis for many scientific problems of interest. The first is the *lasso* regularizer, which encourages overall sparsity in the active (contributing) predictors at each timepoint; the second is the *fused lasso* regularizer, which encourages a notion of persistence or contiguity in the sets of active predictors across timepoints.

Our work is particularly motivated by the analysis of a large data set provided by the Cardiovascular Health Study Cognition Study (CHS-CS). Over the past 24 years, the CHS-CS recorded multiple metabolic, cardiovascular and neuroimaging risk factors for Alzheimer's disease (AD), as well as detailed cognitive assessments for people of ages 65 to 110 years old [Lopez et al., 2003, Saxton et al., 2004, Lopez et al., 2007]. As a matter of background, the prevalence of AD increases at an exponential-like rate beyond the age of 65. After 90 years of age, the incidence of AD increases dramatically, from 12.7% per year in the 90-94 age group, to 21.2% per year in the 95-99 age group, and to 40.7% per year for those older than 100 years [Evans et al., 1989, Fitzpatrick et al., 2004, Corrada et al., 2010]. Later, we examine data from 924 individuals in the Pittsburgh section of the CHS-CS. The objective is to use the data available from subjects at t years of age to predict the onset of AD at $t + 10$ years of age ($\Delta = 10$). For each age, the outcome variable assigns an individual to one of 3 categories: normal, dementia, death. Refer to Section 3 for our analysis of the CHS-CS data set.

*These authors contributed equally to this work

Keywords and phrases: longitudinal data, multinomial logit model, fused lasso, Alzheimer's disease

1.1. *The multinomial fused lasso model.* Given the number of parameters involved in our general longitudinal setup, it will be helpful to be clear about notation: see Table 1. Note that the matrix Y stores future outcome values, i.e., the element Y_{it} records the outcome of the i th individual at time $t + \Delta$, where $\Delta \geq 0$ determines the time lag of the prediction. In the following, we will generally use the “.” symbol to denote partial indexing; examples are $X_{i \cdot t}$, the vector of p predictors for individual i at time t , and $\beta_{\cdot tk}$, the vector of p multinomial coefficients at time t and for class k . Also, Section 2 will introduce an extension of the basic setup in which the number of individuals can vary across timepoints, with n_t denoting the number of individuals at each timepoint $t = 1, \dots, T$.

Parameter	Meaning
$i = 1, \dots, n$	index for individuals
$j = 1, \dots, p$	index for predictors
$t = 1, \dots, T$	index for timepoints
$k = 1, \dots, K$	index for outcomes
Y	$n \times T$ matrix of (future) outcomes
X	$n \times p \times T$ array of predictors
β_0	$T \times (K - 1)$ matrix of intercepts
β	$p \times T \times (K - 1)$ array of coefficients

TABLE 1

Notation used throughout the paper.

At each timepoint $t = 1, \dots, T$, we use a separate multinomial logit model for the outcome at time $t + \Delta$:

$$\begin{aligned}
 (1) \quad & \log \frac{\mathbb{P}(Y_{it} = 1 | X_{i \cdot t} = x)}{\mathbb{P}(Y_{it} = K | X_{i \cdot t} = x)} = \beta_{0t1} + \beta_{\cdot t1}^T x \\
 & \log \frac{\mathbb{P}(Y_{it} = 2 | X_{i \cdot t} = x)}{\mathbb{P}(Y_{it} = K | X_{i \cdot t} = x)} = \beta_{0t2} + \beta_{\cdot t2}^T x \\
 & \vdots \\
 & \log \frac{\mathbb{P}(Y_{it} = K - 1 | X_{i \cdot t} = x)}{\mathbb{P}(Y_{it} = K | X_{i \cdot t} = x)} = \beta_{0t(K-1)} + \beta_{\cdot t(K-1)}^T x.
 \end{aligned}$$

The coefficients are determined by maximizing a penalized log likelihood criterion,

$$(2) \quad (\hat{\beta}_0, \hat{\beta}) \in \arg \max_{\beta_0, \beta} \ell(\beta_0, \beta) - \lambda_1 P_1(\beta) - \lambda_2 P_2(\beta),$$

where $\ell(\beta_0, \beta)$ is the multinomial log likelihood,

$$\ell(\beta_0, \beta) = \sum_{t=1}^T \sum_{i=1}^{n_t} \mathbb{P}(Y_{it} | X_{i \cdot t}),$$

P_1 is the lasso penalty [Tibshirani, 1996],

$$P_1(\beta) = \sum_{j=1}^p \sum_{t=1}^T \sum_{k=1}^{K-1} |\beta_{jtk}|.$$

and P_2 is a version of the fused lasso penalty [Tibshirani et al., 2005] applied across timepoints,

$$P_2(\beta) = \sum_{j=1}^p \sum_{t=1}^{T-1} \sum_{k=1}^{K-1} |\beta_{jtk} - \beta_{j(t+1)k}|.$$

(The element notation in (2) emphasizes the fact that the maximizing coefficients $(\hat{\beta}_0, \hat{\beta})$ need not be unique, since the log likelihood $\ell(\beta_0, \beta)$ need not be strictly concave—e.g., this is the case when $p > n$.)

In broad terms, the lasso and fused lasso penalties encourage sparsity and persistence, respectively, in the estimated coefficients $\hat{\beta}$. A larger value of the tuning parameter $\lambda_1 \geq 0$ generally corresponds to fewer nonzero entries in $\hat{\beta}$; a larger value of the tuning parameter $\lambda_2 \geq 0$ generally corresponds to fewer change points in the piecewise constant coefficient trajectories $\hat{\beta}_{j,k}$, across $t = 1, \dots, T$. We note that the form the log likelihood $\ell(\beta_0, \beta)$ specified above assumes independence between the outcomes across timepoints, which is a rather naive assumption given the longitudinal nature of our problem setup. However, this naivety is partly compensated by the role of the fused lasso penalty, which ties together the multinomial models across timepoints.

It helps to see an example. We consider a simple longitudinal problem with $n = 50$ individuals, $T = 15$ timepoints, and $K = 2$ classes. At each timepoint we sampled $p = 30$ predictors independently from a standard normal distribution. The true (unobserved) coefficient matrix β is now 30×15 ; we set $\beta_j = 0$ for $j = 1 \dots 27$, and set the 3 remaining coefficients trajectories to be piecewise constant across $t = 1, \dots, 15$, as shown in the left panel of Figure 1. In other words, the assumption here is that only 3 of the 30 variables are relevant for predicting the outcome, and these variables have piecewise constant effects over time. We generated a matrix of binary outcomes Y according to the multinomial model (1), and computed the multinomial fused lasso estimates $\hat{\beta}_0, \hat{\beta}$ in (2). The right panel of Figure 1 displays these estimates (all but the intercept $\hat{\beta}_0$) across $t = 1, \dots, 15$, for a favorable choice of tuning parameters $\lambda_1 = 2.5$, $\lambda_2 = 12.5$; the middle plot shows the unregularized (maximum likelihood) estimates corresponding to $\lambda_1 = \lambda_2 = 0$.

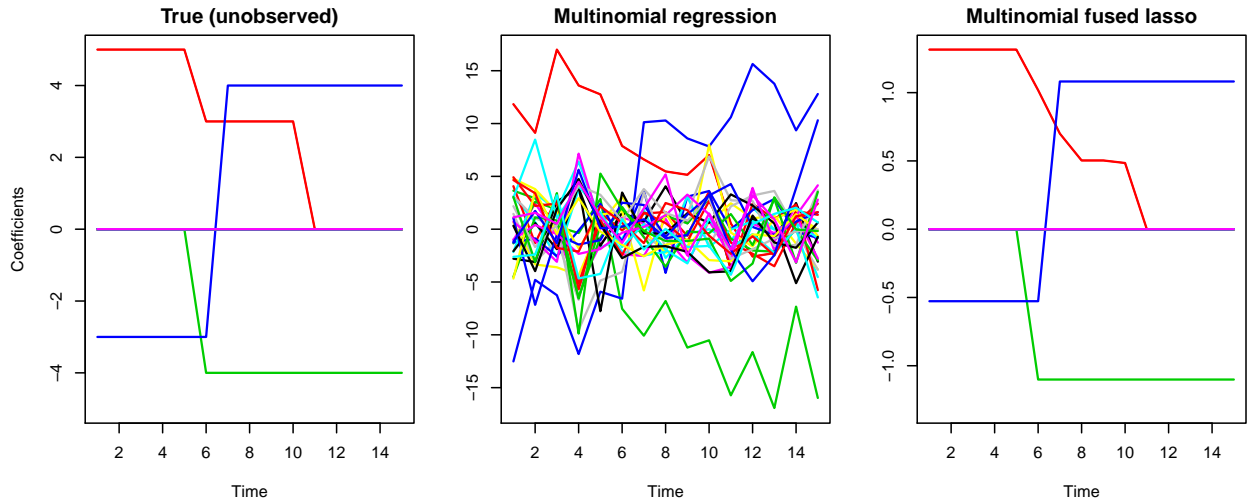


FIG 1. A simple example with $n = 50$, $T = 15$, $K = 2$, and $p = 30$. The left panel displays the true coefficient trajectories across timepoints $t = 1, \dots, 15$ (only 3 of the 30 are nonzero); the middle panel shows the (unregularized) maximum likelihood estimates; the right panel shows the regularized estimates from (1), with $\lambda_1 = 2.5$ and $\lambda_2 = 12.5$.

Each plot in Figure 1 has a y-axis that has been scaled to suit its own dynamic range. We can see that the multinomial fused lasso estimates, with an appropriate amount of regularization, pick up the underlying trend in the true coefficients, though the overall magnitude of coefficients is shrunk toward zero (an expected consequence of the ℓ_1 penalties). In comparison, the unregularized multi-

nomial estimates are wild and do not convey the proper structure. From the perspective of prediction error, the multinomial fused lasso estimates offer a clear advantage, as well: over 30 repetitions from the same simulation setup, we used both the regularized coefficient estimates (with $\lambda_1 = 2.5$ and $\lambda_2 = 12.5$) and the unregularized estimates to predict the outcomes on an i.i.d. test set. The average prediction error using the regularized estimates was 0.114 (with a standard error of 0.014), while the average prediction error from the unregularized estimates was 0.243 (with a standard error of 0.022).

1.2. Related work and alternative approaches. The fused lasso was first introduced in the statistics literature by Tibshirani et al. [2005], and similar ideas based on total variation, starting with Rudin et al. [1992], have been proposed and studied extensively in the signal processing community. There have been many interesting statistical applications of the fused lasso, in problems involving the analysis of comparative genomic hybridization data [Tibshirani and Wang, 2008], the modeling of genome association networks [Kim and Xing, 2009], and the prediction of colorectal cancer [Lin et al., 2013]. The fused lasso has in fact been applied to the study of Alzheimer’s disease in Xin et al. [2014], though these authors consider a very different prediction problem than ours, based on static magnetic resonance images, and do not have the time-varying setup that we do.

Our primary motivation, which is the focus of Section 3, is the problem of predicting the status of an individual at age $t + 10$ years from a number of variables measured at age t . For this we use the regularized multinomial model described in (1), (2). We encode $K = 3$ multinomial categories as normal, dementia, and death: these are the three possible outcomes for any individual at age $t + 10$. We are mainly interested in the prediction of dementia; this task is complicated by the fact that risk factors for dementia are also known to be risk factors for death [Rosvall et al., 2009], and so to account for this, we include the death category in the multinomial classification model. An alternate approach would be to use a Cox proportional hazards model [Cox, 1972], where the event of interest is the onset of dementia, and censorship corresponds to death.

Traditionally, the Cox model is not fit with time-varying predictors or time-varying coefficients, but it can be naturally extended to the setting considered in this work, even using the same regularization schemes. Instead of the multinomial model (1), we would model the hazard function as

$$(3) \quad h(t + \Delta | X_{i:t} = x) = h_0(t + \Delta) \cdot \exp(x^T \beta_t),$$

where $\beta \in \mathbb{R}^{p \times T}$ are a set of coefficients over time, and h_0 is some baseline hazard function (that does not depend on predictor measurements). Note that the hazard model (3) relates the instantaneous rate of failure (onset of dementia) at time $t + \Delta$ to the predictor measurements at time t . This is as in the multinomial model (1), which relates the outcomes at time $t + \Delta$ (dementia or death) to predictor measurements at time t . The coefficients in (3) would be determined by maximizing the partial log likelihood with the analogous lasso and fused lasso penalties on β , as in the above multinomial setting (2).

The partial likelihood approach can be viewed as a sequence of conditional log odds models [Efron, 1977, Kalbfleisch and Prentice, 2002], and therefore one might expect the (penalized) Cox regression model described here to perform similarly to the (penalized) multinomial regression model pursued in this paper. In fact, the computational routine described in Section 2 would apply to the Cox model with only very minor modifications (that concern the gradient computations). A rigorous comparison of the two approaches is beyond the scope of the current manuscript, but is an interesting topic for future development.

1.3. *Outline.* The rest of this paper is organized as follows. In Section 2, we describe a proximal gradient descent algorithm for efficiently computing a solution $(\hat{\beta}_0, \hat{\beta})$ in (1). Next, we present an analysis of the CHS-CS data set in Section 3. Section 4 discusses the stability of estimated coefficients, and related concepts. In Section 5 we discuss numerous approaches for the selecting the tuning parameters $\lambda_1, \lambda_2 \geq 0$ that govern the strength of the lasso and fused lasso penalties in (1). In Section 6, we conclude with some final comments and lay out ideas for future work.

2. A proximal gradient descent approach. In this section, we describe an efficient proximal gradient descent algorithm for computing solutions of the fused lasso regularized multinomial regression problem (2). While a number of other algorithmic approaches are possible, such as implementations of the alternating direction method of multipliers [Boyd et al., 2011], we settle on the proximal gradient method because of its simplicity, and because of the extremely efficient, direct proximal mapping associated with the fused lasso regularizer. We begin by reviewing proximal gradient descent in generality, then we describe its implementation for our problem, and a number of practical considerations like the choice of step size, and stopping criterion.

2.1. *Proximal gradient descent.* Suppose that $g : \mathbb{R}^d \rightarrow \mathbb{R}$ is convex and differentiable, $h : \mathbb{R}^d \rightarrow \mathbb{R}$ is convex, and we are interested in computing a solution

$$x^* \in \operatorname{argmin}_{x \in \mathbb{R}^d} g(x) + h(x).$$

If h were assumed differentiable, then the criterion $f(x) = g(x) + h(x)$ is convex and differentiable, and repeating the simple gradient descent steps

$$(4) \quad x^+ = x - \tau \nabla f(x)$$

suffices to minimize f , for an appropriate choice of step size τ . (In the above, we write x^+ to denote the gradient descent update from the current iterate x .) If h is not differentiable, then gradient descent obviously does not apply, but as long as h is “simple” (to be made precise shortly), we can apply a variant of gradient descent that shares many of its properties, called *proximal gradient descent*. Proximal gradient descent is often also called composite or generalized gradient descent, and in this routine we repeat the steps

$$(5) \quad x^+ = \operatorname{prox}_{h,\tau}(x - \tau \nabla g(x))$$

until convergence, where $\operatorname{prox}_{h,\tau} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is the proximal mapping associated with h (and τ),

$$(6) \quad \operatorname{prox}_{h,\tau}(x) = \operatorname{argmin}_{z \in \mathbb{R}^d} \frac{1}{2\tau} \|x - z\|_2^2 + h(z).$$

(Strict convexity of the above criterion ensures that it has a unique minimizer, so that the proximal mapping is well-defined.) Provided that h is simple, by which we mean that its proximal map (6) is explicitly computable, the proximal gradient descent steps (5) are straightforward and resemble the classical gradient descent analogues (4); we simply take a gradient step in the direction governed by the smooth part g , and then apply the proximal map of h . A slightly more formal perspective argues that the updates (6) are the result of minimizing h plus a quadratic expansion of g , around the current iterate x .

Proximal gradient descent has become a very popular tool for optimization problems in statistics and machine learning, where typically g represents a smooth loss function, and h a nonsmooth regularizer. This trend is somewhat recent, even though the study of proximal mappings has a long

history of in the optimization community (e.g., see [Parikh and Boyd \[2013\]](#) for a nice review paper). In terms of convergence properties, proximal gradient descent enjoys essentially the same convergence rates as gradient descent under the analogous assumptions, and is amenable to acceleration techniques just like gradient descent (e.g., [Nesterov \[2007\]](#), [Beck and Teboulle \[2009\]](#)). Of course, for proximal gradient descent to be applicable in practice, one must be able to exactly (or even approximately) compute the proximal map of h in (6); fortunately, this is possible for many optimization problems, i.e., many common regularizers h , that are encountered in statistics. In our case, the proximal mapping reduces to solving a problem of the form

$$(7) \quad \hat{\theta} = \argmin_{\theta} \frac{1}{2} \|x - \theta\|_2^2 + \lambda_1 \sum_{i=1}^m |\theta_i| + \lambda_2 \sum_{i=1}^{m-1} |\theta_i - \theta_{i+1}|.$$

This is often called the fused lasso signal approximator (FLSA) problem, and extremely fast, linear-time algorithms exist to compute its solution. In particular, we rely on an elegant dynamic programming approach proposed by [Johnson \[2013\]](#).

2.2. Application to the multinomial fused lasso problem. The problem in (2) fits into the desired form for proximal gradient descent, with g the multinomial regression loss (i.e., negative multinomial regression log likelihood) and h the lasso plus fused lasso penalties. Formally, we can rewrite (2) as

$$(8) \quad (\hat{\beta}_0, \hat{\beta}) \in \argmin_{\beta_0, \beta} g(\beta_0, \beta) + h(\beta_0, \beta),$$

where g is the convex, smooth function

$$g(\beta_0, \beta) = \sum_{t=1}^T \sum_{i=1}^n \left\{ \sum_{k=1}^{K-1} -\mathbb{I}(Y_{it} = k)(\beta_{0tk} + X_{i \cdot t} \beta_{\cdot tk}) + \log \left(1 + \sum_{h=1}^{K-1} \exp(\beta_{0th} + X_{i \cdot t} \beta_{\cdot th}) \right) \right\},$$

and h is the convex, nonsmooth function

$$h(\beta_0, \beta) = \lambda_1 \sum_{j=1}^p \sum_{t=1}^T \sum_{k=1}^{K-1} |\beta_{jtk}| + \lambda_2 \sum_{j=1}^p \sum_{t=1}^{T-1} \sum_{k=1}^{K-1} |\beta_{jtk} - \beta_{j(t+1)k}|.$$

Here we consider fixed values $\lambda_1, \lambda_2 \geq 0$. As described previously, each of these tuning parameters will have a big influence on the strength of their respective penalty terms, and hence the properties of the computed estimate $(\hat{\beta}_0, \hat{\beta})$; we discuss the selection of λ_1 and λ_2 in Section 5. We note that the intercept coefficients β_0 are not penalized.

To compute the proximal gradient updates, as given in (5), we must consider two quantities: the gradient of g , and the proximal map of h . First, we discuss the gradient. As $\beta_0 \in \mathbb{R}^{T \times (K-1)}$, $\beta \in \mathbb{R}^{p \times T \times (K-1)}$, we may consider the gradient as having dimension $\nabla g(\beta_0, \beta) \in \mathbb{R}^{(p+1) \times T \times (K-1)}$. We will index this as $[\nabla g(\beta_0, \beta)]_{jtk}$ for $j = 0, \dots, p$, $t = 1, \dots, T$, $k = 1, \dots, K-1$; hence note that $[\nabla g(\beta_0, \beta)]_{0tk}$ gives the partial derivative of g with respect to β_{0tk} , and $[\nabla g(\beta_0, \beta)]_{jtk}$ the partial derivative with respect to β_{jtk} , for $j = 1, \dots, p$. For generic t, k , we have

$$(9) \quad [\nabla g(\beta_0, \beta)]_{0tk} = \sum_{i=1}^n \left(-\mathbb{I}(Y_{it} = k) + \frac{\exp(\beta_{0tk} + X_{i \cdot t} \beta_{\cdot tk})}{1 + \sum_{h=1}^{K-1} \exp(\beta_{0th} + X_{i \cdot t} \beta_{\cdot th})} \right),$$

and for $j \geq 1$,

$$(10) \quad [\nabla g(\beta_0, \beta)]_{jtk} = \sum_{i=1}^n \left(-\mathbb{I}(Y_{it} = k) X_{ijt} + X_{ijt} \frac{\exp(\beta_{0tk} + X_{i \cdot t} \beta_{\cdot tk})}{1 + \sum_{h=1}^{K-1} \exp(\beta_{0th} + X_{i \cdot t} \beta_{\cdot th})} \right).$$

It is evident that computation of the gradient requires $O(npTK)$.

Now, we discuss the proximal operator. Since the intercept coefficients $\beta_0 \in \mathbb{R}^{T \times (K-1)}$ are left unpenalized, the proximal map over β_0 just reduces to the identity, and the intercept terms undergo the updates

$$\beta_{0tk}^+ = \beta_{0tk} - \tau[\nabla g(\beta_0, \beta)]_{0tk} \text{ for } t = 1, \dots, T, k = 1, \dots, K-1.$$

Hence we consider the proximal map over β alone. At an arbitrary input $x \in \mathbb{R}^{p \times T \times (K-1)}$, this is

$$\underset{z \in \mathbb{R}^{p \times T \times (K-1)}}{\operatorname{argmin}} \frac{1}{2\tau} \sum_{j=1}^p \sum_{t=1}^T \sum_{k=1}^{K-1} (x_{jtk} - z_{jtk})^2 + \lambda_1 \sum_{j=1}^p \sum_{t=1}^T \sum_{k=1}^{K-1} |\beta_{jtk}| + \lambda_2 \sum_{j=1}^p \sum_{t=1}^{T-1} \sum_{k=1}^{K-1} |\beta_{jtk} - \beta_{j(t+1)k}|,$$

which we can see decouples into $p(K-1)$ separate minimizations, one for each predictor $j = 1, \dots, p$ and class $k = 1, \dots, K-1$. In other words, the coefficients β undergo the updates

$$(11) \quad \beta_{j \cdot k}^+ = \underset{\theta \in \mathbb{R}^T}{\operatorname{argmin}} \frac{1}{2} \sum_{t=1}^T \left((\beta_{j \cdot k} - \tau[\nabla g(\beta_0, \beta)]_{j \cdot k}) - \theta \right)^2 + \tau \lambda_1 \sum_{t=1}^T |\theta_t| + \tau \lambda_2 \sum_{t=1}^{T-1} |\theta_t - \theta_{t+1}|,$$

for $j = 1, \dots, p, k = 1, \dots, K-1$,

each minimization being a fused lasso signal approximator problem [Tibshirani et al., 2005], i.e., of the form (7). There are many computational approaches that may be applied to such a problem structure; we employ a specialized, highly efficient algorithm by Johnson [2013] that is based on dynamic programming. This algorithm requires $O(T)$ operations for each of the problems in (11), making the total cost of the update $O(pTK)$ operations. Note that this is actually dwarfed by the cost of computing the gradient $\nabla g(\beta_0, \beta)$ in the first place, and therefore the total complexity of a single iteration of our proposed proximal gradient descent algorithm is $O(npTK)$.

2.3. Practical considerations. We discuss several practical issues that arise in applying the proximal gradient descent algorithm.

2.3.1. Backtracking line search. Returning to the generic perspective for proximal gradient descent as described in Section 2.1, we rewrite the proximal gradient descent update in (5) as

$$(12) \quad x^+ = x - \tau G_\tau(x),$$

where $G_\tau(x)$ is called the *generalized gradient* and is defined as

$$G_\tau = \frac{x - \operatorname{prox}_{h, \tau}(x - \tau \nabla g(x))}{\tau}.$$

The update is rewritten in this way so that it more closely resembles the usual gradient update in (4). We can see that, analogous to the gradient descent case, the choice of parameter $\tau > 0$ in each iteration of proximal gradient descent determines the magnitude of the update in the direction of the generalized gradient $G_\tau(x)$. Classical analysis shows that if ∇g is Lipschitz with constant $L > 0$, then proximal gradient descent converges with any fixed choice of step size $\tau \leq 1/L$ across all iterations. In most practical situations, however, the Lipschitz constant L of ∇g is not known or easily computable, and we rely on an adaptive scheme for choosing an appropriate step size at each iteration; backtracking line search is one such scheme, which is straightforward to implement in practice and guarantees convergence of the algorithm under the same Lipschitz assumption on ∇g (but importantly, without having to know its Lipschitz constant L). Given a shrinkage factor

$0 < \gamma < 1$, the backtracking line search routine at a given iteration of proximal gradient descent starts with $\tau = \tau_0$ (a large initial guess for the step size), and while

$$(13) \quad g(x - \tau G_\tau(x)) > g(x) - \tau \nabla g(x)^T G_\tau(x) + \frac{\tau}{2} \|G_\tau(x)\|_2^2,$$

it shrinks the step size by letting $\tau = \gamma\tau$. Once the exit criterion is achieved (i.e., the above is no longer satisfied), the proximal gradient descent algorithm then uses the current value of τ to take an update step, as in (12) (or (5)).

In the case of the multinomial fused lasso problem, the generalized gradient is of dimension $G_\tau(\beta_0, \beta) \in \mathbb{R}^{(p+1) \times T \times (K-1)}$, where

$$[G_\tau(\beta_0, \beta)]_{0..} = [\nabla g(\beta_0, \beta)]_{0..},$$

and

$$[G_\tau(\beta_0, \beta)]_{j..k} = \frac{\beta_{j..k} - \text{prox}_{\text{FLSA}, \tau}(\beta_{j..k} - \tau [\nabla g(\beta_0, \beta)]_{j..k})}{\tau} \quad \text{for } j = 1, \dots, p, k = 1, \dots, K-1.$$

Here $\text{prox}_{\text{FLSA}, \tau}(\beta_{j..k} - \tau [\nabla g(\beta_0, \beta)]_{j..k})$ is the proximal map defined by the fused lasso signal approximator evaluated at $\beta_{j..k} - \tau [\nabla g(\beta_0, \beta)]_{j..k}$, i.e., the right-hand side in (11). Backtracking line search now applies just as described above.

2.3.2. Stopping criteria. The simplest implementation of proximal gradient descent would run the algorithm for a fixed, large number of steps S . A more refined approach would check a stopping criterion at the end of each step, and terminate if such a criterion is met. Given a tolerance level $\epsilon > 0$, two common stopping criteria are then based on the relative difference in function values, as in

$$\text{stopping criterion 1: terminate if } C_1 = \frac{|f(\beta_0^+, \beta^+) - f(\beta_0, \beta)|}{f(\beta_0, \beta)} \leq \epsilon,$$

and the relative difference in iterates, as in

$$\text{stopping criterion 2: terminate if } C_2 = \frac{\|(\beta_0^+, \beta^+) - (\beta_0, \beta)\|_2}{\|(\beta_0, \beta)\|_2} \leq \epsilon.$$

The second stopping criterion is generally more stringent, and may be hard to meet in large problems, given a small tolerance ϵ .

For the sake of completeness, we outline the full proximal gradient descent procedure in the notation of the multinomial fused lasso problem, with backtracking line search and the first stopping criterion, in Algorithms 1 and 2 below.

2.3.3. Missing individuals. Often in practice, some individuals are not present at some time-points in the longitudinal study, meaning that one or both of their outcome values and predictor measurements are missing over a subset of $t = 1, \dots, T$. Let I_t denote the set of completely observed individuals (i.e., with both predictor measurements and outcomes observed) at time t , and let $n_t = |I_t|$. The simplest strategy to accomodate such missingness would be to compute the loss function g only observed individuals, so that

$$g(\beta_0, \beta) = \sum_{t=1}^T \sum_{i \in I_t} \left\{ \sum_{k=1}^{K-1} -\mathbb{I}(Y_{it} = k)(\beta_{0tk} + X_{i..t} \beta_{..tk}) + \log \left(1 + \sum_{h=1}^{K-1} \exp(\beta_{0th} + X_{i..t} \beta_{..th}) \right) \right\}.$$

Algorithm 1 Proximal gradient descent for the multinomial fused lasso

INPUT: Predictors X , outcomes Y , tuning parameter values λ_1, λ_2 , initial coefficient guesses $(\beta_0^{(0)}, \beta^{(0)})$, maximum number of iterations S , initial step size before backtracking τ_0 , backtracking shrinkage parameter γ , tolerance ϵ

OUTPUT: Approximate solution $(\hat{\beta}_0, \hat{\beta})$

```
1:  $s = 1, C = \infty$ 
2: while ( $s \leq S$  and  $C > \epsilon$ ) do
3:   Find  $\tau_s$  using backtracking, Algorithm 2 (INPUT:  $\beta_0^{(s-1)}, \beta^{(s-1)}, \tau_0, \gamma$ )
4:   Update the intercept:  $\beta_{0..}^{(s)} = \beta_{0..}^{(s-1)} - \tau_s [\nabla g(\beta_0^{(s-1)}, \beta^{(s-1)})]_{0..}$ 
5:   for  $j = 1, \dots, p$  do
6:     for  $k = 1, \dots, (K-1)$  do
7:       Update  $\beta_{j:k}^{(s)} = \text{prox}_{\text{FLSA}, \tau_s}(\beta_{j:k}^{(s-1)} - \tau_s [\nabla g(\beta_0^{(s-1)}, \beta^{(s-1)})]_{j:k})$ 
8:     end for
9:   end for
10:  Increment  $s = s + 1$ 
11:  Compute  $C = [f(\beta_0^{(s)}, \beta^{(s)}) - f(\beta_0^{(s-1)}, \beta^{(s-1)})] / f(\beta_0^{(s-1)}, \beta^{(s-1)})$ 
12: end while
13:  $\hat{\beta}_0 = \beta_0^{(s)}, \hat{\beta} = \beta^{(s)}$ 
14: return  $(\beta_0, \hat{\beta})$ 
```

Algorithm 2 Backtracking line search for the multinomial fused lasso

INPUT: $\beta_0, \beta, \tau_0, \gamma$

OUTPUT: τ

```
1:  $\tau = \tau_0$ 
2: while (true) do
3:   Compute  $[G_\tau(\beta_0, \beta)]_{0..} = [\nabla g(\beta_0, \beta)]_{0..}$ 
4:   for  $j = 1, \dots, p$  do
5:     for  $k = 1, \dots, (K-1)$  do
6:       Compute  $[G_\tau(\beta_0, \beta)]_{j:k} = [\beta_{j:k} - \text{prox}_{\text{FLSA}, \tau}(\beta_{j:k} - \tau [\nabla g(\beta_0, \beta)]_{j:k})] / \tau$ 
7:     end for
8:   end for
9:   if  $g((\beta_0, \beta) - \tau G_\tau(\beta_0, \beta)) > g(\beta_0, \beta) - \tau [\nabla g(\beta_0, \beta)]^T G_\tau(\beta_0, \beta) + \frac{\tau}{2} \|G_\tau(\beta_0, \beta)\|_2^2$  then
10:    Break
11:   else
12:    Shrink  $\tau = \gamma \tau$ 
13:   end if
14: end while
15: return  $\tau$ 
```

An issue arises when the effective sample size n_t is quite variable across timepoints t : in this case, the penalty terms can have quite different effects on the coefficients $\beta_{..t}$ at one time t versus another. That is, the coefficients $\beta_{..t}$ at a time t in which n_t is small experience a relatively small loss term

$$(14) \quad \sum_{i \in I_t} \left\{ \sum_{k=1}^{K-1} -\mathbb{I}(Y_{it} = k)(\beta_{0tk} + X_{i \cdot t} \beta_{\cdot tk}) + \log \left(1 + \sum_{h=1}^{K-1} \exp(\beta_{0th} + X_{i \cdot t} \beta_{\cdot th}) \right) \right\},$$

simply because there are fewer terms in the above sum compared to a time with a larger effective sample size; however, the penalty term

$$\lambda_1 \sum_{j=1}^p \sum_{k=1}^{K-1} |\beta_{jtk}| + \lambda_2 \sum_{j=1}^p \sum_{k=1}^{K-1} |\beta_{jtk} - \beta_{j(t+1)k}|$$

remains comparable across all timepoints, regardless of sample size. A fix would be to scale the loss term in (14) by n_t to make it (roughly) independent of the effective sample size, so that the total loss

becomes

$$(15) \quad g(\beta_0, \beta) = \sum_{t=1}^T \frac{1}{n_t} \sum_{i \in I_t} \left\{ \sum_{k=1}^{K-1} -\mathbb{I}(Y_{it} = k)(\beta_{0tk} + X_{i \cdot t} \beta_{\cdot tk}) + \log \left(1 + \sum_{h=1}^{K-1} \exp(\beta_{0th} + X_{i \cdot t} \beta_{\cdot th}) \right) \right\}.$$

This modification indeed ends up being important for the Alzheimer’s analysis that we present in Section 3, since this study has a number of individuals in the tens at some timepoints, and in the hundreds for others. The proximal gradient descent algorithm described in this section extends to cover the loss in (15) with only trivial modifications.

2.4. Implementation in C++ and R. An efficient C++ implementation of the proximal gradient descent algorithm described in this section, with an easy interface to R, is available from the second author’s website: <http://www.stat.cmu.edu/~flecci>. In the future, this will be available as part of the R package `glmgen`, which broadly fits generalized linear models under generalized lasso regularization.

3. Alzheimer’s Disease data analysis. In this section, we apply the proposed estimation method to the data of the the Cardiovascular Health Study Cognition Study (CHS-CS), a rich database of thousands of multiple cognitive, metabolic, cardiovascular, cerebrovascular, and neuroimaging variables obtained over the past 24 years for people of ages 65 to 110 years old [Fried et al., 1991, Lopez et al., 2007].

The complex relationships between age and other risk factors produce highly variable natural histories from normal cognition to the clinical expression of Alzheimer’s disease, either as dementia or its prodromal syndrome, mild cognitive impairment (MCI) [Lopez et al., 2003, Saxton et al., 2004, Lopez et al., 2007, Sweet et al., 2012, Lecci, 2014]. Many studies involving the CHS-CS data have shown the importance of a range of risk factors in predicting the time of onset of clinical dementia. The risk of dementia is affected by the presence of the APOE*4 allele, male sex, lower education, and having a family history of dementia [Fitzpatrick et al., 2004, Tang et al., 1996, Launer et al., 1999]. Medical risks include the presence of systemic hypertension, diabetes mellitus, and cardiovascular or cerebrovascular disease [Kuller et al., 2003, Irie et al., 2005, Skoog et al., 1996]. Lifestyle factors affecting risk include physical and cognitive activity, and diet [Verghese et al., 2003, Erickson et al., 2010, Scarmeas et al., 2006].

A wide range of statistical approaches has been considered in these studies, including exploratory statistical summaries, hypothesis tests, survival analyses, logistic regression models, and latent trajectory models. None of these methods can directly accommodate a large number of predictors that can potentially exceed the number of observations. A small number of variables was often chosen a priori to match the requirements of a particular model, neglecting the full potential of the CHS-CS data, which consists of thousands of variables.

The approach that we introduced in Section 1 can accommodate an array of predictors of arbitrary dimension, using regularization to maintain a well-defined predictive model and avoid overfitting. Our goal is to identify important risk factors for the prediction of the cognitive status at $t + 10$ years of age ($\Delta = 10$), given predictor measurements at t years of age, for $t = 65, 66, \dots, 98$. We use the penalized log likelihood criterion in (2) to estimate the coefficients of the multinomial logit model in (1). The lasso penalty forces the solution to be sparse, allowing us to identify a few important predictors among the thousands of variables of the CHS-CS data. The fused lasso penalty allows for a few change points in the piecewise constant coefficient trajectories $\hat{\beta}_{j \cdot k}$, across $t = 65, \dots, 98$. Justification for this second penalty is based on the scientific intuition that predictors that are clinically important should have similar effects in successive ages.

3.1. *Data preprocessing.* We use data from the $n = 924$ individuals in the Pittsburgh section of the CHS-CS, recorded between 1990 and 2012. Each individual underwent clinical and cognitive assessments at multiple ages, all falling in the range 65, ... 108. The matrix of (future) outcomes Y has dimension $n \times 34$: for $i = 1, \dots, 924$ and $t = 65, \dots, 98$, the outcome Y_{it} stores the cognitive status at age $t + 10$ and can assume one of the following values:

$$Y_{it} = \begin{cases} 1 & \text{if normal} \\ 2 & \text{if MCI/dementia .} \\ 3 & \text{if dead} \end{cases}$$

MCI is included in the same class as dementia, as they are both instances of cognitive impairment. Hence the proposed multinomial model predicts the onset of MCI/dementia, in the presence of a separate death category. This is done to implicitly adjust for the confounding effect of death, as some risk factors for dementia are also known to be risk factors for death [Rosvall et al., 2009].

The array of predictors X is composed of time-varying variables that were recorded at least twice during the CHS-CS study, and time-invariant variables, such as gender and race. A complication in the data set is the ample amount of missingness in the array of predictors. We impute missing values using a uniform rule for all possible causes of missingness. A missing value at age t is imputed by taking the closest past measurement from the same individual, if present. If all the past values are missing, the global median from people of age t is used. The only exception is the case of time-invariant predictors, whose missing values are imputed by either future or past values, as available.

Categorical variables with m possible outcomes are converted to $m - 1$ binary variables and all the predictors are standardized to have zero mean and unit standard deviation. This is a standard procedure in regularization, as the lasso and fused lasso penalties puts constraints on the size of the coefficients associated with each variable [Tibshirani, 1997]. To be precise, imputation of missing values and standardization of the predictors are performed within each of the folds used in the cross-validation method for the choice of the tuning parameters λ_1 and λ_2 (discussed below), and then again for the full data set in the final estimation procedure that uses the selected tuning parameters.

The final array of predictors X has dimension $924 \times 1050 \times 34$, where 1050 is the number of variables recorded over the period of 34 years of age range.

3.2. *Model and algorithm specification.* In the Alzheimer’s Disease application, the multinomial model in (1) is determined by two equations, as there are three possible outcomes (normal, MCI/dementia, death); the outcome “normal” is taken as the base class. We will refer to the two equations (and the corresponding sets of coefficients) as the “dementia vs normal” and “death vs normal” equations, respectively.

We use the proximal gradient descent algorithm described in Section 2 to estimate the coefficients that maximize the penalized log likelihood criterion in (2). The initializations $(\beta_0^{(0)}, \beta^{(0)})$ are set to be zero matrices, the maximum number of iterations is $S = 80$, the initial step size before backtracking is $\tau_0 = 20$, the backtracking shrinkage parameter is $\gamma = 0.6$ and the tolerance of the first stopping criterion (relative difference in function values) is $\epsilon = 0.001$. We select the tuning parameters by a 4-fold cross-validation procedure that minimizes the misclassification error. The selected parameters are $\lambda_1 = 0.019$ and $\lambda_2 = 0.072$, which yield an average prediction error of 0.316 (standard error 0.009). Section 5 discusses more details on the model selection problem.

The number n_t of outcomes observed at age t varies across time, for two reasons: first, different subjects entered the study at different ages, and second, once a subject dies at time t_0 , we exclude them consideration in the model formed at all ages $t > t_0$, to predict the outcomes of individuals at

age $t + 10$. The maximum number of outcomes is 604 at age 88, whereas the minimum is 7 at age 108. We resort to the strategy described in Section 2.3.3 and use the scaled loss in (15) to compensate for the varying sample sizes.

3.3. *Results.* Out of the 1050 coefficients associated with the predictors described above, 148 are estimated to be nonzero for at least one time point in the 34 years age range. More precisely, for at least one age, 57 coefficients are nonzero in the “dementia vs normal” equation of the predictive multinomial logit model, and 124 are nonzero in the “death vs normal” equation.

Dementia vs normal	
Variable	Meaning (and coding for categorical variables, before scaling)
race01.2	Race: "White" 1, else 0
cdays59	Taken vitamin C in the last 2 weeks? (number of days)
newthg68.1	How is the person at learning new things wrt 10 yrs ago? "A bit worse" 1, else 0
estrop39	If you not currently taking estrogen, have you taken in the past? "Yes" 1, "No" 0
fear05.1	How often felt fearful during last week? "Most of the time" 1, else 0
early39	Do you usually wake up far too early? "Yes" 1, "No" 0
gend01	Gender: "Female" 1, "Male" 0
hctz06	Medication: thiazide diuretics w/o K-sparing. "Yes" 1, "No" 0
race01.1	Race: "Other (no white, no black)" 1, else 0
orthos27	Do you use a lower extremity orthosis? "Yes" 1, "No" 0
pulse21	60 second heart rate
grpsym09.1	What causes difficulty in gripping? "Pain in arm/hand" 1, else 0
sick03.2	If sick, could easily find someone to help? "Probably False" 1, else 0
digcor	Digit-symbol substitution task: number of symbols correctly coded
trust03.3	There is at least one person whose advice you really trust. "Probably true" 1, else 0

Death vs normal	
Variable	Meaning (and coding for categorical variables, before scaling)
digcor	Digit-symbol substitution task: number of symbols correctly coded
ctime27	Repeated chair stands: number of seconds
gend01	Gender: "Female" 1, "Male" 0
cis42	Cardiac injury score
hurry59.2	Ever had pain in chest when walking uphill/hurry? "No" 1, else 0
numcig59	Number of cigarettes smoked per day
dig06	Digitalis medicines prescribed? "Yes" 1, "No" 0
smoke.3	Current smoke status: "Never smoked" 1, else 0
hlth159.1	Would you say, in general, your health is.. ? "Fair" 1, else 0
exer59	If gained/lost weight, was exercise a major factor? "Yes" 1, "No" 0
nomeds06	Number of medications taken
diabada.3	ADA diabetic status? "New diabetes" 1, else 0
anyone	Does anyone living with you smoke cigarettes regularly? "Yes" 1, "No" 0
ltaai	Blood pressure variable: left ankle-arm index
whmile09.2	Do you have difficulty walking one-half a mile? "Yes" 1, else 0

TABLE 2

The 15 most important variables in the two separate equations of the multinomial logit model.

Figure 2 shows the 15 most important variables in the 34 years age range, separately for the two equations. The measure of importance is described in detail in Section 4 and is, in fact, a measure of stability of the estimated coefficients, across 4 subsets of the data (the 4 training sets used in cross-validation). The plots on the left show the relative importance of the 15 variables with respect to the most important one, whose importance was scaled to be 100. The plots on the right show, separately for the two equations, the longitudinal estimated coefficients for the 15 most important variables,

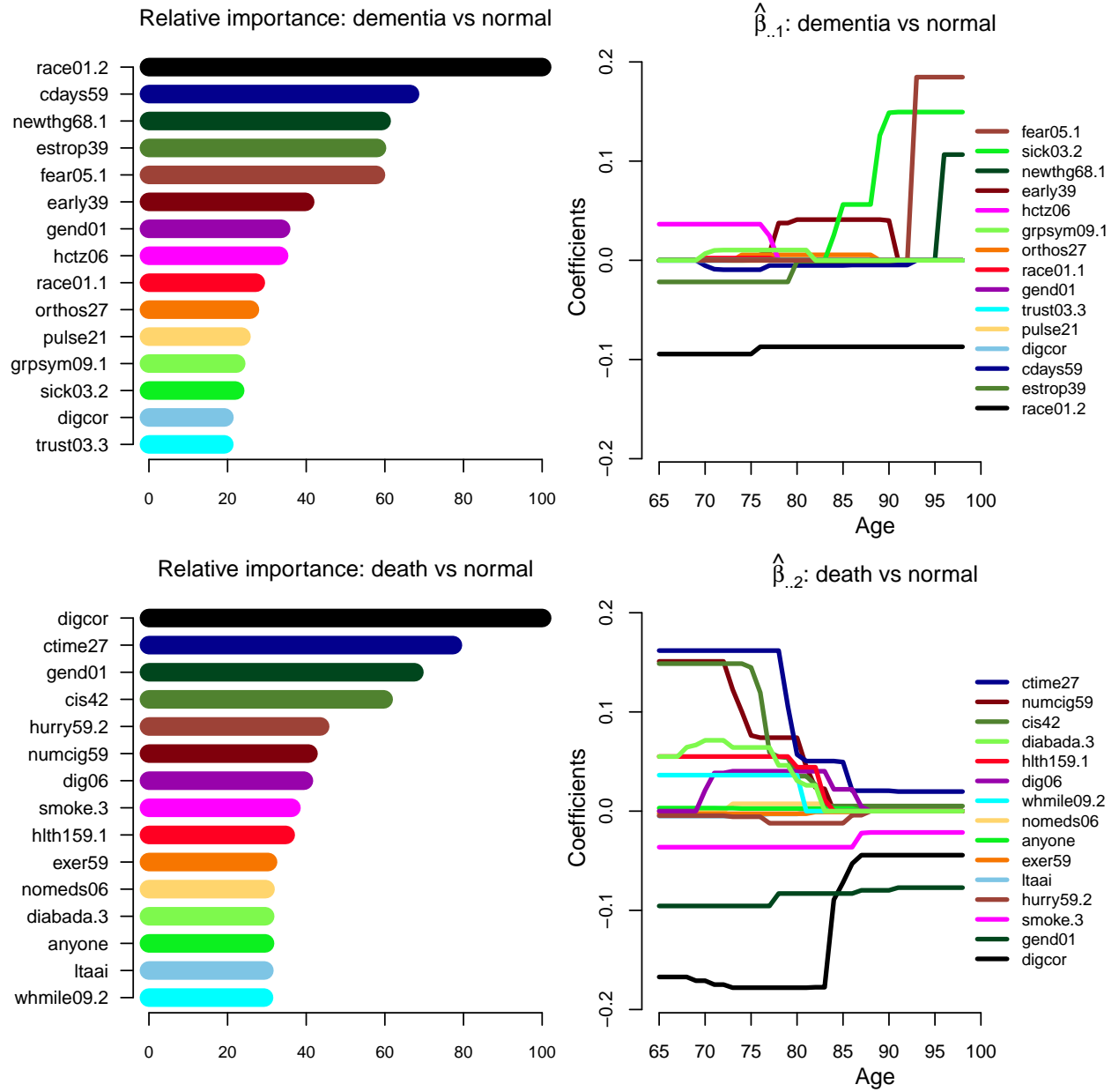


FIG 2. CHS-CS data analysis. Left: relative importance plots for the 15 most important variables in the “dementia vs normal” and “death vs normal” equations of the multinomial logit model. Right: corresponding estimated coefficients. The order of the legends follow the order of the maximum/minimum values of the estimated coefficient trajectories. Note that some coefficients are estimated to be very close to 0 and the corresponding trajectories are hidden by other coefficients.

using the data and algorithm specification described above. The meaning of these predictors and the coding used for the categorical variables are reported in Table 2. The nonzero coefficients that are not displayed in Figure 2 are less important (according to our measure of stability) and, for the vast majority, their absolute values are less than 0.1.

We now proceed to interpret the results, keeping in mind that, ultimately, we are estimating the

coefficients of a multinomial logit model and that the outcome variable is recorded 10 years in the future with respect to the predictors. For example, an increase in the value of a predictor with positive estimated coefficient in the top right plot of Figure 2 is associated with an increase of the (10 years future) odds of dementia with respect to a normal cognitive status. In what follows, to facilitate the exposition of results, our statements are less formal.

Inspecting the “dementia vs normal” plot we see that, in general, being Caucasian (race01.2) is associated with a decrease in the odds of dementia, while, after the age of 85, fear (fear05.1), lack of available caretakers (sick03.2), and deterioration of learning skills (newthg68.1) increase the odds of dementia. Variables hctz06 (a particular diuretic) and early39 (early wake-ups) have positive coefficients for the age ranges 65,...78 and 77,...91, respectively, and hence, if active, they account for an increase of the risk of dementia. The “death vs normal” plot reveals the importance of several variables in the age range 65,...85: longer time to rise from sitting in a chair (ctime27), more cigarettes (numcig59), higher cardiac injury score (cis42) are associated with an increase of the odds of death. Other variables in the same age range, with analogous interpretations, but lower importance, are diabada.3 (“new diabetes” diagnosis), hlth159.1 (“fair” health status), dig06 (use of Digitalis), whmile09.2 (difficulty in walking). By contrast, in the same age range, good performance on the digit-symbol substitution task (digcor) accounts for a decrease in the odds of death. Finally, regardless of the age, being a non-smoker (smoke.3) or being a woman (gend01) decrease the odds of death.

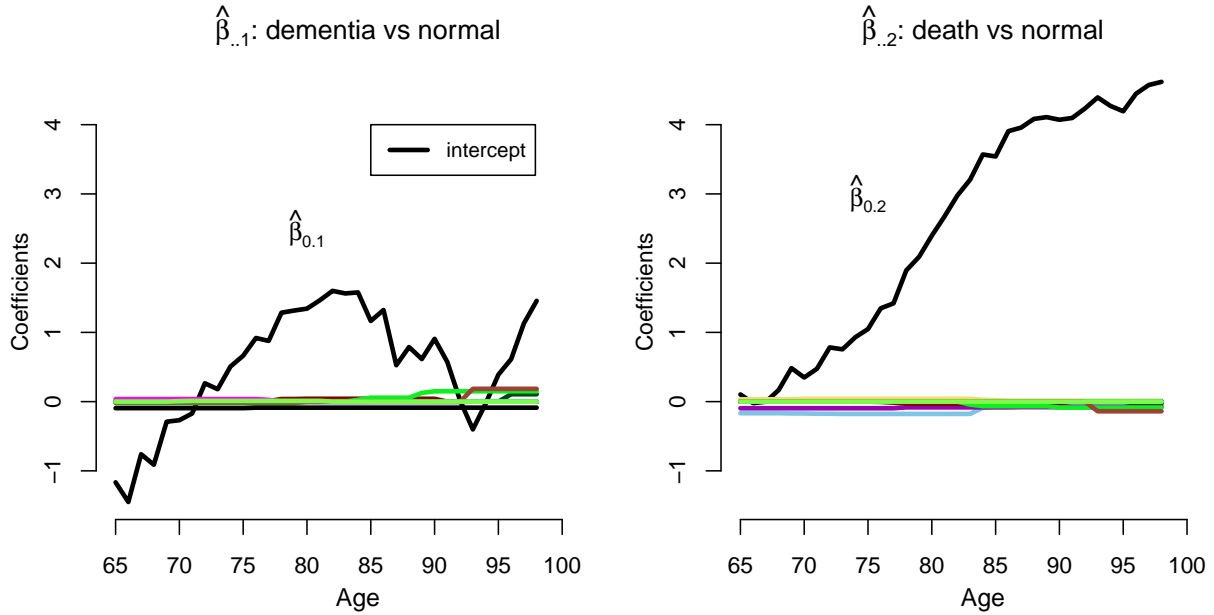


FIG 3. CHS-CS data analysis. Estimated intercept coefficients in the two separate equations of the multinomial logit model.

Figure 3 shows the intercept coefficients $\hat{\beta}_{0,1}$ and $\hat{\beta}_{0,2}$, which, we recall, are not penalized in the log likelihood criterion in (2). The intercepts account for time-varying risk that is not explained by the predictors. In particular, the coefficients $\hat{\beta}_{0,2}$ increases over time, suggesting that an increasing amount of risk of death can be attributed to a subject’s age alone, independent of the predictor measurements.

3.4. Discussion. The results of the proposed multinomial fused lasso methodology applied to the CHS data are broadly consistent with what is known about risk and protective factors for dementia in the elderly [Lopez et al., 2013]. Race, gender, vascular and heart disease, lack of available caregivers, and deterioration of learning and memory are all associated with an increased risk of dementia. The results, however, provide critical new insights into the natural progression of MCI/dementia. First, the relative importance of the risk factors changes over time. As shown in Figure 2, with the exception of race, risk factors for dementia become more relevant after the age of 85. This is critical, as there is increasing evidence [Kuller et al., 2011] for a change in the risk profile for the expression of clinical dementia among the oldest-old. Second, the independent prediction of death, and the associated risk/protection factors, highlight the close connection between risk of death and risk of dementia. That is, performance on a simple, timed test of psychomotor speed (digit symbol substitution task) is a very powerful predictor of death within 10 years, as is a measure of physical strength/frailty (time to arise from a chair). Other variables, including gender, diabetes, walking and exercise, are all predictors of death, but are known, from other analyses in the CHS and other studies, to be linked to the risk of dementia. The importance of these risk/protective factors for death is attenuated (with the exception of gender) after age 85, likely reflecting survivor bias. Taken together, these results add to the growing body of evidence of the critical importance of accounting for mortality in the analysis of risk for dementia, especially among the oldest old [Kuller et al., 2011].

For our analysis we chose a 10 year time window for risk prediction. Among individuals of age 65-75, who are cognitively normal, this may be a scientifically and clinically reasonable time window to use. However, had we similar data from individuals as young as 45-50 years old, then we might wish to choose time windows of 20 years or longer. In the present case, it could be argued that a shorter time window might be more scientifically and clinically relevant among individuals over the age of 80 years, as survival times of 10 years become increasingly less likely in the oldest-old.

4. Measures of stability. Examining the stability of variables in a fitted model, subject to small perturbations of the data set, is one way to assess variable importance. Applications of stability, in this spirit, have recently gained popularity in the literature, across a variety of settings such as clustering (e.g., Lange et al. [2004]), regression (e.g., Meinshausen and Bühlmann [2010]), and graphical models (e.g., Liu et al. [2010]). Here we propose a very simple stability-based measure of variable importance, based on the definition of variable importance for trees and additive tree expansions [Breiman et al., 1984, Hastie et al., 2008]. We fit the multinomial fused lasso estimate (2) on the data set $X_{i\cdot}, Y_{i\cdot}$, for $i = i_1, \dots, i_m$, a subsample of the total individuals $1, \dots, n$, and repeat this process R times. Let $\hat{\beta}^{(r)}$ denote the coefficients from the r th subsampled data set, for $r = 1, \dots, R$. Then we define the importance of variable j for class k as

$$(16) \quad I_{jk} = \frac{1}{RT} \sum_{r=1}^R \sum_{t=1}^T |\hat{\beta}_{jtk}^{(r)}|,$$

for each $j = 1, \dots, p$ and $k = 1, \dots, K - 1$, which is the average absolute magnitude of the coefficients for the j th variable and k th class, across all timepoints, and subsampled data sets. Therefore, a larger value of I_{jk} indicates a higher variable importance, as measured by stability (not only across subsampled data sets r , but actually across timepoints t , as well). Relative importances can be computed by scaling the highest variable importance to be 100, and adjusting the other values accordingly; for simplicity we typically consider relative variable importances in favor of absolute ones, because the original scale has no real meaning.

There is some subtlety in the role of the tuning parameters λ_1, λ_2 used to fit the coefficients $\hat{\beta}^{(r)}$ on each subsampled data set $r = 1, \dots, R$. Note that the importance measure (16) reflects the impor-

tance of a variable in the context of a fitting procedure that, given data samples, produces estimated coefficients. The simplest approach would be to consider the fitting procedure defined by the multinomial fused lasso problem (2) at a fixed pair of tuning parameter values λ_1, λ_2 . But in practice, it is seldom true that appropriate tuning parameter values are known ahead of time, and one typically employs a method like cross-validation to select parameter values (see Section 5 for a discussion of cross-validation and other model selection methods). Hence in this case, to determine variable importances in the final coefficient estimates, we would take care to define our fitting procedure in (16) to be the one that, given data samples, performs cross-validation on these data samples to determine the best choice of λ_1, λ_2 , and then uses this choice to fit coefficient estimates. In other words, for each subsampled data set $r = 1, \dots, R$ in (16), we would perform cross-validation to determine tuning parameter values and then compute $\hat{\beta}^{(r)}$ as the multinomial fused lasso solution at these chosen parameter values. This is more computationally demanding, but it is a more accurate reflection of variable importance in the final model output by the multinomial fused lasso under cross-validation for model selection.

The relative variable importances for the CHS-CS data example from Section 3 are displayed in Figure 2, alongside the plots of estimated coefficients. Here we drew 4 subsampled data sets, each one containing 75% of the total number of individuals. The tuning parameter values have been selected by cross-validation. The variable importances were defined to incorporate this selection step into the fitting procedure, as explained above. We observe that the variables with high positive or negative coefficients for most ages in the plotted trajectories typically also have among the highest relative importances. Another interesting observation concerns categorical predictors, which (recall) have been converted into binary predictors over multiple levels: often only some levels of a categorical predictor are active in the plotted trajectories.

5. Model selection. The selection of tuning parameters λ_1, λ_2 is clearly an important issue that we have not yet covered. In this section, we discuss various methods for automatic tuning parameter selection in the multinomial fused lasso model (2), and apply them to a subset of the CHS-CS study data of Section 3, with 140 predictors and 600 randomly selected individuals, as an illustration. In particular, we consider the following methods for model selection: cross-validation, cross-validation under the one-standard-error rule, AIC, BIC, and finally AIC and BIC using misclassification loss (in place of the usual negative log likelihood). Note that cross-validation in our longitudinal setting is performed by dividing the individuals $1, \dots, n$ into folds, and, per its typical usage, selecting the tuning parameter pair λ_1, λ_2 (over, say, a grid of possible values) that minimizes the cross-validation misclassification loss. The one-standard-error rule, on the other hand, picks the simplest estimate that achieves a cross-validation misclassification loss within one standard error of the minimum. Here “simplest” is interpreted to mean the estimate with the fewest number of nonzero component blocks. AIC and BIC scores are computed for a candidate λ_1, λ_2 pair by

$$\begin{aligned} \text{AIC}(\lambda_1, \lambda_2) &= 2 \cdot \text{loss}((\hat{\beta}_0, \hat{\beta})_{\lambda_1, \lambda_2}) + 2 \cdot \text{df}((\hat{\beta}_0, \hat{\beta})_{\lambda_1, \lambda_2}), \\ \text{BIC}(\lambda_1, \lambda_2) &= 2 \cdot \text{loss}((\hat{\beta}_0, \hat{\beta})_{\lambda_1, \lambda_2}) + \log N_{\text{tot}} \cdot \text{df}((\hat{\beta}_0, \hat{\beta})_{\lambda_1, \lambda_2}), \end{aligned}$$

and in each case, the tuning parameter pair is chosen (again, say, over a grid of possible values) to minimize the score. In the above, $(\hat{\beta}_0, \hat{\beta})_{\lambda_1, \lambda_2}$ denotes the multinomial fused lasso estimate (2) at the tuning parameter pair λ_1, λ_2 , and N_{tot} denotes the total number of observations in the longitudinal study, $N_{\text{tot}} = nT$ (or $N_{\text{tot}} = \sum_{t=1}^T n_t$ in the missing data setting). Also, $\text{df}((\hat{\beta}_0, \hat{\beta})_{\lambda_1, \lambda_2})$ denotes the degrees of freedom of the estimate $(\hat{\beta}_0, \hat{\beta})_{\lambda_1, \lambda_2}$, and we employ the approximation

$$\text{df}((\hat{\beta}_0, \hat{\beta})_{\lambda_1, \lambda_2}) \approx \# \text{ of nonzero blocks in } (\hat{\beta}_0, \hat{\beta})_{\lambda_1, \lambda_2},$$

borrowing from known results in the Gaussian likelihood case [Tibshirani and Taylor, 2011, 2012]. Finally, $\text{loss}((\hat{\beta}_0, \hat{\beta})_{\lambda_1, \lambda_2})$ denotes a loss function considered for the estimate, which we take either to be the negative multinomial log likelihood $-\ell((\hat{\beta}_0, \hat{\beta})_{\lambda_1, \lambda_2})$, as is typical in AIC and BIC [Hastie et al., 2008], or the misclassification loss, to put it on closer footing to cross-validation. Note that both loss functions are computed in-sample, i.e., over the training samples, and hence AIC and BIC are computationally much cheaper than cross-validation.

We compare these model selection methods on the subset of the CHS-CS data set. The individuals are randomly split into 5 folds. We use 4/5 of the data set to perform model selection and subsequent model fitting with the 6 techniques described above: cross-validation, cross-validation with the one-standard-error rule, and AIC and BIC under negative log likelihood and misclassification losses. To be perfectly clear, the model selection techniques work entirely within this given 4/5 of the data set, so that, e.g., cross-validation further divides this data set into folds. In fact, we used 4-fold cross-validation to make this division simplest. The remaining 1/5 of the data set is then used for evaluation of the estimates coming from each of the 6 methods, and this entire process is repeated, leaving out each fold in turn as the evaluation set. We record several measures on each evaluation set: the misclassification rate, true positive rate in identifying the dementia class, true positive rate in identifying the dementia and death classes combined, and degrees of freedom (number of nonzero blocks in the estimate). Figure 4 displays the mean and standard errors of these 4 measures, for each of the 6 model selection methods.

Cross-validation and cross-validation with the one-standard-error rule both seem to represent a favorable balance between the different evaluation measures. The cross-validation methods provide a misclassification rate significantly better than that of the null model, which predicts according to the majority class (death), they yield two of the three highest true positive rates in identifying the dementia class, and perform well in terms of identifying the dementia and death classes combined (as do all methods: note that all true positive rates here are about 0.75 or higher). We ended up settling cross-validation under the usual rule, rather than the one-standard-error rule, because the former achieves the highest true positive rate in identifying the dementia class, which was our primary concern in the CHS-CS data analysis. By design, cross-validation with the one-standard-error rule delivers a simpler estimate in terms of degrees of freedom (196 for the one-standard-error rule versus 388 for the usual rule) though both cross-validation models are highly regularized in absolute terms (e.g., the fully saturated model would have thousands of nonzero blocks).

6. Discussion and future work. In this work, we proposed a multinomial model for high-dimensional longitudinal classification tasks. Our proposal operates under the assumption that a sparse number of predictors contribute more or less persistent effects across time. The multinomial model is fit under lasso and fused lasso regularization, which address the assumptions of sparsity and persistence, respectively, and lead to piecewise constant estimated coefficient profiles. We described a highly efficient computational algorithm for this model based on proximal gradient descent, demonstrated the applicability of this model on an Alzheimer’s data set taken from the CHS-CS, and discussed practically important issues such stability measures for the estimates and tuning parameter selection.

A number of extensions of the basic model are well within reach. For example, placing a group lasso penalty on the coefficients associated with each level of a binary expansion for a categorical variable may be useful for encouraging sparsity in a group sense (i.e., over all levels of a categorical variable at once). As another example, more complex trends than piecewise constant ones may be fit by replacing the fused lasso penalty with a trend filtering penalty [Kim et al., 2009, Tibshirani, 2014], which would lead to piecewise polynomial trends of any chosen order k . The appropriateness

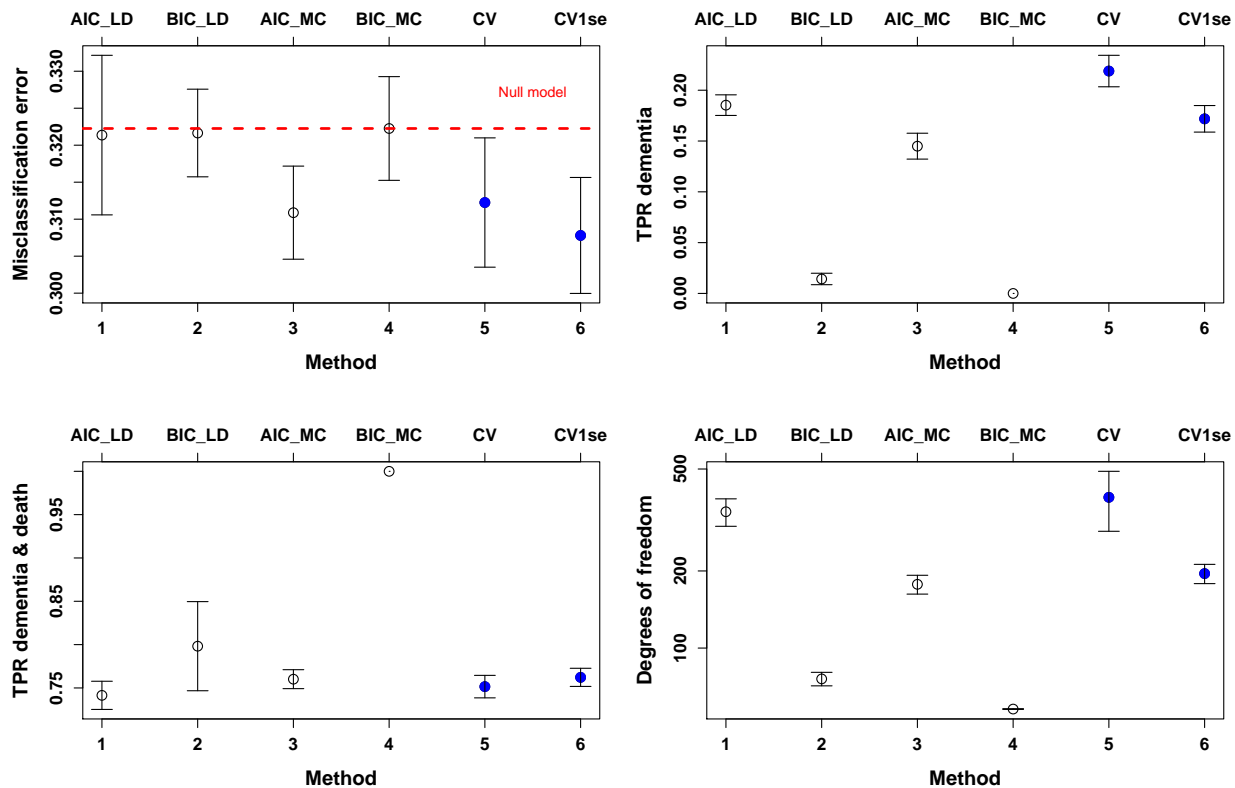


FIG 4. Comparison of different methods for selection of tuning parameters λ_1, λ_2 on the CHS-CS data set. The x-axis in each plot parametrizes the 6 different methods considered, which are, from left to right: AIC and BIC under negative log likelihood loss, AIC and BIC under misclassification loss, cross-validation, and cross-validation with the one-standard-error rule. The upper left plot shows (out-of-sample) misclassification rate associated with the estimates selected by each method, averaged over 5 iterations. The segments denote ± 1 standard errors around the mean. The red dotted line is the average misclassification rate associated with the naive estimator that predicts all individuals as dead (the majority class). The upper right and bottom left plots show different measures of evaluation (again, computed out-of-sample): the true positive rate in identifying the dementia class, respectively, the true positive rate in identifying the dementia and death classes combined. Finally, the bottom right plot shows degrees of freedom (number of nonzero blocks) of the estimates selected by each method.

of such a penalty would depend on the scientific application; the use of a fused lasso penalty assumes that the effect of a given variable is mostly constant across time, with possible change points; the use of a quadratic trend filtering penalty (polynomial order $k = 2$) allows the effect to vary more smoothly across time.

More difficult and open-ended extensions concern statistical inference for the fitted longitudinal classification models. For example, the construction of confidence intervals (or bands) for selected coefficients (or coefficient profiles) would be an extremely useful tool for the practitioner, and would offer more concrete and rigorous interpretations than the stability measures described in Section 4. Unfortunately, this is quite a difficult problem, even for simpler regularization schemes (such as a pure lasso penalty) and simpler observation models (such as linear regression). But recent inferential developments for related high-dimensional estimation tasks [Zhang and Zhang, 2011, Javanmard and Montanari, 2013, van de Geer et al., 2013, Lockhart et al., 2014, Lee et al., 2013, Taylor et al.,

2014] shed a positive light on this future endeavor.

References.

- Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.
- Steve Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternative direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.
- Leo Breiman, Jerome Friedman, Charles Stone, and Richard Olshen. *Classification and Regression Trees*. Chapman & Hall/CRC Press, Boca Raton, 1984.
- Maria M Corrada, Ron Brookmeyer, Annlia Paganini-Hill, Daniel Berlau, and Claudia H Kawas. Dementia incidence continues to increase with age in the oldest old: the 90+ study. *Annals of neurology*, 67(1):114–121, 2010.
- David Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, 34(2):187–220, 1972.
- Bradley Efron. The efficiency of Cox’s likelihood function for censored data. *Journal of the American Statistical Association*, 72(359):557–565, 1977.
- KI Erickson, CA Raji, OL Lopez, JT Becker, C Rosano, AB Newman, HM Gach, PM Thompson, AJ Ho, and LH Kuller. Physical activity predicts gray matter volume in late adulthood the cardiovascular health study. *Neurology*, 75(16):1415–1422, 2010.
- Denis A Evans, H Harris Funkenstein, Marilyn S Albert, Paul A Scherr, Nancy R Cook, Marilyn J Chown, Liesi E Hebert, Charles H Hennekens, and James O Taylor. Prevalence of alzheimer’s disease in a community population of older persons. *JAMA: the journal of the American Medical Association*, 262(18):2551–2556, 1989.
- Annette L Fitzpatrick, Lewis H Kuller, Diane G Ives, Oscar L Lopez, William Jagust, John Breitner, Beverly Jones, Constantine Lyketsos, and Corinne Dulberg. Incidence and prevalence of dementia in the cardiovascular health study. *Journal of the American Geriatrics Society*, 52(2):195–204, 2004.
- Linda P Fried, Nemat O Borhani, Paul Enright, Curt D Furberg, Julius M Gardin, Richard A Kronmal, Lewis H Kuller, Teri A Manolio, Maurice B Mittelman, Anne Newman, et al. The cardiovascular health study: design and rationale. *Annals of epidemiology*, 1(3):263–276, 1991.
- Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning; Data Mining, Inference and Prediction*. Springer, New York, 2008. Second edition.
- F Irie, AL Fitzpatrick, OL Lopez, R Peila, LH Kuller, A Newman, and LJ Launer. Type 2 diabetes (t2d), genetic susceptibility and the incidence of dementia in the cardiovascular health study. In *Neurology*, volume 64, pages A316–A316, 2005.
- Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. arXiv: 1306.3171, 2013.
- Nicholas Johnson. A dynamic programming algorithm for the fused lasso and L_0 -segmentation. *Journal of Computational and Graphical Statistics*, 22(2):246–260, 2013.
- John Kalbflesich and Ross Prentice. *The Statistical Analysis of Failure Time Data*. Wiley Series in Probability and Statistics, New Jersey, 2002.
- Seung-Jean Kim, Kwangmoo Koh, Stephen Boyd, and Dimitry Gorinevsky. ℓ_1 trend filtering. *SIAM Review*, 51(2):339–360, 2009.
- Seyoung Kim and Eric Xing. Statistical estimation of correlated genome associations to a quantitative trait network. *PLoS Genetics*, 5(8):e1000587, 08 2009.
- Lewis Kuller, Yuefang Chang, James Becker, and Oscar Lopez. Does alzheimer’s disease over 80 years old have a different etiology? *Alzheimer’s & Dementia*, 7(4):S596–S597, 2011.
- Lewis H Kuller, Oscar L Lopez, Anne Newman, Norman J Beauchamp, Greg Burke, Corinne Dulberg, Annette Fitzpatrick, Linda Fried, and Mary N Haan. Risk factors for dementia in the cardiovascular health cognition study. *Neuroepidemiology*, 22(1):13–22, 2003.
- Tilman Lange, Volker Roth, Mikio Braun, and Joachim Buhmann. Stability-based validation of clustering solutions. *Neural Computation*, 16:1299–1323, 2004.
- LJ Launer, K Andersen, MEea Dewey, L Letenneur, A Ott, LA Amaducci, C Brayne, JRM Copeland, J-F Dartigues, P Kragh-Sorensen, et al. Rates and risk factors for dementia and alzheimer’s disease results from eurodem pooled analyses. *Neurology*, 52(1):78–84, 1999.
- Fabrizio Lecci. An analysis of development of dementia through the Extended Trajectory Grade of Membership model. In Edoardo M. Airoldi, David Blei, Elena A. Erosheva, and Stephen E. Fienberg, editors, *Handbook of Mixed Membership Models and Their Applications*. Chapman & Hall/CRC, 2014.
- Jason Lee, Dennis Sun, Yukai Sun, and Jonathan Taylor. Exact post-selection inference with the lasso. arXiv: 1311.6238, 2013.

- Yunzhi Lin, Menggang Yu, Sijian Wang, Richard Chappell, , and Thomas Imperiale. Advanced colorectal neoplasia risk stratification by penalized logistic regression. *Statistical Methods in Medical Research*, 2013.
- Han Liu, Kathryn Roeder, and Larry Wasserman. Stability approach to regularization selection (StARS) for high-dimensional graphical models. *Neural Information Processing Systems*, 23, 2010.
- Richard Lockhart, Jonathan Taylor, Ryan J. Tibshirani, and Robert Tibshirani. A significance test for the lasso. *Annals of Statistics*, 42(2):413–468, 2014.
- Oscar L Lopez, William J Jagust, Steven T DeKosky, James T Becker, Annette Fitzpatrick, Corinne Dulberg, John Breitner, Constantine Lyketsos, Beverly Jones, Claudia Kawas, et al. Prevalence and classification of mild cognitive impairment in the cardiovascular health study cognition study: part 1. *Archives of neurology*, 60(10):1385, 2003.
- Oscar L Lopez, Lewis H Kuller, James T Becker, Corinne Dulberg, Robert A Sweet, H Michael Gach, and Steven T DeKosky. Incidence of dementia in mild cognitive impairment in the cardiovascular health study cognition study. *Archives of Neurology*, 64(3):416–420, 2007.
- Oscar L Lopez, James T Becker, and Lewis H Kuller. Patterns of compensation and vulnerability in normal subjects at risk of alzheimer’s disease. *Journal of Alzheimer’s Disease*, 33:S427–S438, 2013.
- Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B*, 72(4): 417–473, 2010.
- Yurii Nesterov. Gradient methods for minimizing composite objective function. CORE discussion paper, 2007.
- Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):123–231, 2013.
- Lina Rosvall, Debora Rizzuto, Hui-Xin Wang, Bengt Winblad, Caroline Graff, and Laura Fratiglioni. Apoe-related mortality: Effect of dementia, cardiovascular disease and gender. *Neurobiology of Aging*, 30(10):1545–1551, 2009.
- Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60:259–268, 1992.
- J Saxton, Oscar L Lopez, G Ratcliff, C Dulberg, LP Fried, MC Carlson, AB Newman, and L Kuller. Preclinical alzheimer disease neuropsychological test performance 1.5 to 8 years prior to onset. *Neurology*, 63(12):2341–2347, 2004.
- Nikolaos Scarmeas, Yaakov Stern, Ming-Xin Tang, Richard Mayeux, and Jose A Luchsinger. Mediterranean diet and risk for Alzheimer’s disease. *Annals of Neurology*, 59(6):912–921, 2006.
- I Skoog, L Nilsson, G Persson, B Lernfelt, S Landahl, B Palmertz, LA Andreasson, A Oden, and A Svanborg. 15-year longitudinal study of blood pressure and dementia. *The Lancet*, 347(9009):1141–1145, 1996.
- Robert A Sweet, Howard Seltman, James E Emanuel, Oscar L Lopez, James T Becker, Joshua C Bis, Elise A Weamer, Mary Ann A DeMichele-Sweet, and Lewis H Kuller. Effect of alzheimer’s disease risk genes on trajectories of cognitive function in the cardiovascular health study. *American Journal of Psychiatry*, 169(9):954–962, 2012.
- Ming-Xin Tang, Gladys Maestre, Wei-Yann Tsai, Xin-Hua Liu, Lin Feng, Wai-Yee Chung, Michael Chun, Peter Schofield, Yaakov Stern, Benjamin Tycko, et al. Relative risk of alzheimer disease and age-at-onset distributions, based on apoe genotypes among elderly african americans, caucasians, and hispanics in new york city. *American journal of human genetics*, 58(3):574–584, 1996.
- Jonathan Taylor, Richard Lockhart, Ryan J. Tibshirani, and Robert Tibshirani. Exact post-selection inference for forward stepwise and least angle regression. arXiv: 1401.3889, 2014.
- Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288, 1996.
- Robert Tibshirani. The lasso method for variable selection in the cox model. *Statistics in Medicine*, 16(4):385–395, 1997.
- Robert Tibshirani and Pei Wang. Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics*, 9(1):18–29, January 2008.
- Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B*, 67(1):91–108, 2005.
- Ryan J. Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. *Annals of Statistics*, 42(1):285–323, 2014.
- Ryan J. Tibshirani and Jonathan Taylor. The solution path of the generalized lasso. *Annals of Statistics*, 39(3):1335–1371, 2011.
- Ryan J. Tibshirani and Jonathan Taylor. Degrees of freedom in lasso problems. *Annals of Statistics*, 40(2):1198–1232, 2012.
- Sara van de Geer, Peter Bühlmann, and Ya’acov Ritov. On asymptotically optimal confidence regions and tests for high-dimensional models. arXiv: 1303.0518, 2013.
- Joe Verghese, Richard B Lipton, Mindy J Katz, Charles B Hall, Carol A Derby, Gail Kuslansky, Anne F Ambrose, Martin Sliwinski, and Herman Buschke. Leisure activities and the risk of dementia in the elderly. *New England Journal of Medicine*, 348(25):2508–2516, 2003.
- Bo Xin, Yoshinobu Kawahara, Yizhou Wang, and Wen Gao. Efficient generalized fused lasso and its application to the diagnosis of alzheimer’s disease. *AAAI Conference on Artificial Intelligence*, 28, 2014.
- Cun-Hui Zhang and Stephanie Zhang. Confidence intervals for low-dimensional parameters with high-dimensional data.

arXiv: 1110.2563, 2011.

S. ADHIKARI
F. LECCI
B.W. JUNKER
R.J. TIBSHIRANI
DEPARTMENT OF STATISTICS
CARNEGIE MELLON UNIVERSITY
PITTSBURGH, PA 15213
E-MAIL: asamrach@andrew.cmu.edu
lecci@cmu.edu
brian@stat.cmu.edu
ryantibs@cmu.edu

L.H. KULLER
DEPARTMENT OF EPIDEMIOLOGY
UNIVERSITY OF PITTSBURGH
PITTSBURGH, PA 15213
E-MAIL: kullerl@edc.pitt.edu

J.T. BECKER
O.L. LOPEZ
DEPARTMENT OF NEUROLOGY
UNIVERSITY OF PITTSBURGH
PITTSBURGH, PA 15213
E-MAIL: beckerJT@upmc.edu
lopezOL@upmc.edu